Optical-flow kymograms and glottovibrograms: A new way to present high-speed data for laryngeal assessment

Conference Paper · September 2015

CITATIONS
0

READS
97

3 authors, including:

Gustavo Andrade-Miranda
Universidad Politécnica de Madrid
13 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Nathalie Henrich Bernardoni
French National Centre for Scientific Research
136 PUBLICATIONS **1,122** CITATIONS
SEE PROFILE

OPTICAL-FLOW KYMOGRAMS AND GLOTTOVIBROGRAMS: A NEW WAY TO PRESENT HIGH-SPEED DATA FOR LARYNGEAL ASSESSMENT

Gustavo Andrade-Miranda[1], Nathalie Henrich Bernardoni [2,3], Juan Ignacio Godino-Llorente[1]

[1] Center for Biomedical Technologies, Universidad Politécnica de Madrid
[2] Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
[3] CNRS, GIPSA-Lab, F-38000 Grenoble, France
gxandrade@ics.upm.es, Nathalie.Henrich@gipsa-lab.fr, igodino@ics.upm.es

Abstract: **The use of high-speed videoendoscopy (HSV) in combination with image-processing tech- niques is the most promising approach to investigate vocal-folds vibration and laryngeal dynamics in speech and singing. The current challenge is to pro- vide facilitative and informative playbacks for clinical and research purposes. We present three new facilita- tive playbacks using an optical-flow framework (OF), which has the main advantage of requiring no glottis segmentation. The method has been tested on a data- base of 60 HSV sequences, which covers different voice qualities for spoken and sung vowels. The new data representations have been compared with com- monly used facilitative playbacks.**

Keywords: **Optical Flow, motion field, high-speed videoendoscopy, glottal dynamics, playbacks**

I. INTRODUCTION

Nowadays, high-speed videondoscopy has been increasingly used to assess glottal dynamics. It is the sole imaging technique capable to acquire the true intra-cycle vibratory behavior through a series of full-frame images of the vocal folds. It allows the study of cycle-to-cycle glottal variation. Due to the fast-growth of high-speed technology, it is possible to found cameras that can reach frame rates up to "twenty thousands", recording in color with high spatial resolution and excellent image quality for long durations. HSV allows to characterize many vocal-folds vibratory features that are not possible to visualize by means of videostroboscopic techniques. For instance, HSV helps to get insights into tissue vibratory characteristics, the influence of aerodynamical forces and muscular tension, vocal length and evaluation of normal laryngeal functioning in situation of rapid pitch change such as onset and offset of voicing or glides [1]. The use of HSV has been reported in the literature to evaluate variations in vocal-folds dynamics and extract important features such as vocal-fold vibratory amplitude, glottal open quotient, and glottal speed quotient. HSV provides a huge amount of images, whose analysis requires a great deal of human intervention and observation. Several playbacks have been proposed to reduce the spatio-temporal dimensionality while preserving the most relevant characteristics of glottal vibratory patterns. The most widespread and successful playbacks used either by clinicians or researchers are: Digital Kymograms [2], Mucosal Wave and Mucosal Wave Kymogram playbacks [3], Phonovibrogram [4], and Glottovibrogram [5]. Many common approaches make use of glottal segmentation algorithms, in which attention is focused on analyzing movements at vocal-fold edges. The widespread techniques are based on histogram equalization, region growing, watershed and active contours delineation methods (see [5] for a review). Nevertheless, motion analysis should not necessarily be focused only on the points belonging to the glottis contours but also in the regions where such movements originate. For that reason, the estimation of a global motion in which the different patterns relevant for voice production could be represented is desirable. Optical flow (OF) techniques estimate the motion of objects in consecutive frames by generating a motion field in which each pixel represents a vector displacement. In laryngeal HSV sequences, the vibrating vocal folds are most often the regions with greatest motion.

In this paper, OF image processing is investigated as a new approach for analyzing laryngeal images and for assessing glottal dynamics. Innovative playbacks are proposed within this framework. The paper is organized as follows. Section 2 details the principles of OF-based image processing, and describes the data- base. Section 3 presents the results for three new play- backs and provides a comparison with existing com- mon ones. Finally, Section 4 presents some conclu- sions and discussions.

II. MATERIALS AND METHODS

A. Principles of optical-flow estimation

The 3-D velocity vector of objects, projected onto the image plane, is known as the image flow field. This could be considered as the ideal and actual movement of objects that we expect to see. Unfortunately, image processing has to deal with the inverse problem: the movement of the objects has to be determined on the

Claudia Manfredi (edited by), *Models and analysis of vocal emissions for biomedical applications : 9 th international workshop : September 2-4, 2015*, ISBN 978-88-6655-792-0 (print) ISBN 978-88-6655-793-7 (online)

basis of a sequence of images. This leads to an approximation called optical flow field, which associate each pixel in the image with a motion vector.
There are many different ways to estimate the optical flow, which depend basically on the kind of chosen constraint. Most of the constraints are derived from the assumption that pixel intensities are translated from one frame to the next:

$$f(x+\otimes x, y+\otimes y, t+\otimes t) \approx f(x,y,t) \quad (1)$$

Another type of constraint that has no obvious connection with the previous one is motion tensor (MT) [8]. The MT principle is that a video segment is a stack of images in which gray-value structures have certain orientations. The orientation in the *xy*-subspace is an indicator for the orientation of the structure in the space. In contrast, the orientation of the structure in the *xt*-subspace or *yt*-subspace relates to the image velocities. Thus, estimating the orientation of the structure in these two subspaces or a combination thereof allows estimating the OF. There are two main strategies for solving the OF problem: Sparse and Dense. The sparse optical flow finds the displacement only on a subset of features that have been specified beforehand; these features have certain desirable properties such as corners, dominant gradient orientation, or subpixel corner locations. In the other hand, dense OF finds out the vector displacement of all pixels in the image, requiring a more expensive computational burden, but providing more interesting information about the movements in the sequence.

B. Database

A database of 60 high-speed sequences was used to assess laryngeal dynamics in several phonatory tasks: spoken vowels with specific voice qualities (creaky, normal, breathy, pressed), pitch glides, sung vowels at different pitches, loudness and laryngeal mechanisms [6]. Two male subjects (one speaker, one singer) participated to the experiment. The recording took place at the University Medical Center Hamburg-Eppendorf (UKE) in Germany, in collaboration with Pr. Hess, Dr. Müller and Dr. Licht [5]. The high-speed sequences were acquired by means of Wolf high-speed cinematographic system (rigid endoscope Wolf 90 E 60491 and light source Wolf 5131, grayscale CCD camera). The laryngeal high-speed images were sampled at either 2000 or 4000 fps. They had a spatial resolution of 256x256 pixels. Audio and electroglottographic signals were recorded simultaneously to the high-speed sequences and synchronized in a post-processing step.

C. Image Processing Procedure

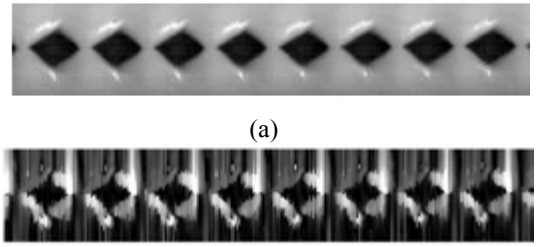
The algorithms were developed in C++ using the OpenCV library and integrated in Matlab for making the visualization of playbacks easier. Two different optical-flow methods were used. The first one, called TV-L1 OF, is based on the brightness constancy assumption [7]. This formulation adds a regularization term that allows discontinuities. Such feature is desirable when a complex motion is modeling. The brightness constancy term uses the robust L1 norm and is therefore less sensitive to intensity variations. The second OF method, called MT OF, is based on motion tensor computation [8]. It starts with computing 3D orientation tensors from the image sequence. These tensors are combined under the constraints of a parametric motion model to produce the velocity estimation. The formulation of this OF methodological approach does not use the common brightness constraint, and thus it is more sensible to the reflectance phenomena originated by the mucosa surface properties. Many additional techniques can be applied to mitigate these effects. The approach chosen here combines a nonlinear transformation with an anisotropic filter. Taking into account that the analysis of laryngeal HSV focuses attention on the dynamics of vocal-folds movements, a good strategy to reduce computational burden and mitigate the effect produced by noise regions is to calculate the OF field inside of a region of interest (ROI) that include the glottal gap and part of the vocal folds. The next step is to synthesize the motion-field information obtained between consecutive frames into visual playbacks, in which the information on the behavior of vocal-folds movement is readable. Three main representations have been elaborated. They will now be described and compared to the existing playbacks.

III. RESULTS

A. Local dynamics along one line: Optical-Flow Kymogram

The Optical-Flow Kymogram playback (OFKG) uses the same principle than Digital Kymogram (DKG) to compact high-speed information. However, the information used to condense the data is taken from the displacements originated in the x-axis. For rightwise OF movements, the direction angle of displacement ranges from $[-\pi/2, \pi/2]$ and is coded in white. On the other hand, the direction angle for leftwise displacements ranges from $[\pi/2, 3\pi/2]$ and is coded in gray tone of 128. The OFKG playback is illustrated in Fig.1 for a sequence of eight glottal cycles. Glottal-cycle shape and glottal dynamics present great similarity in both playbacks. The instants of change between opening and closing phases induce the presence of a discontinuity in the OFKG. This can be understood as the instants for which velocity comes close to zero. The

spread effect in the OFKG at given moments of the opening and closing phases may reflect the mucosal waves on vocal-folds surface. In DKG, mucosal waves are reflected as white flashing spots.



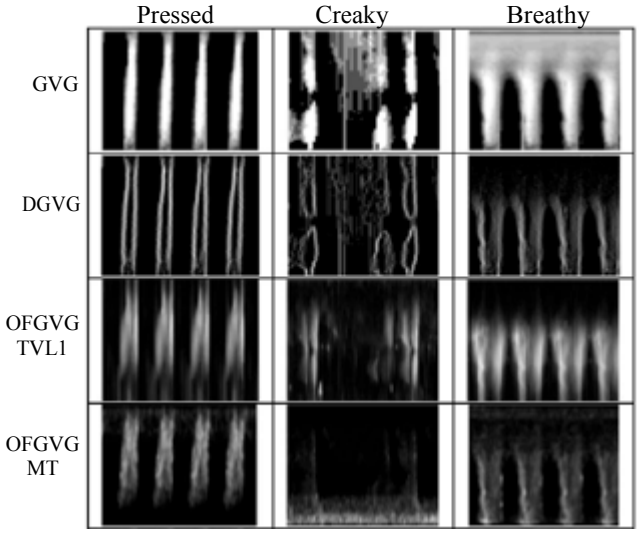
(a)

(b)

Fig.1: (a) DKG representation for a line located in the center of the main axis; (b) The new OFKG playback for the same line, in which gray scale distinguishes the direction of motion (rightwise: white gray; leftwise: pale gray).

B. Global dynamics along the whole vocal-folds length: Optical-Flow Glottovibrogram

The Optical-Flow Glottovibrogram (OFGVG) represents the velocity of glottal movement per cycle plotted along the vocal-folds length. It is obtained by averaging each row of the x component of the flow and representing it as a column vector. This procedure is repeated along time for each new frame. The aim of OFGVG playback is to complement the spatio-temporal information provided by the common techniques (glottovibrogram GVG, phonovibrogram PVG), by adding velocity information for each displacement of the vocal folds. For the purpose of visual comparison, four playbacks were performed in different phonation cases: GVG and its derivative DGVG were computed using [5]; two OFGVG playbacks were computed using TVL1 OF (OFGVG-TVL1) and MT OF (OFGVG-MT) respectively. Only OFGVG-MT includes the preprocessing step described in the section 2C. The corresponding plots for these playbacks are presented in Fig.2.



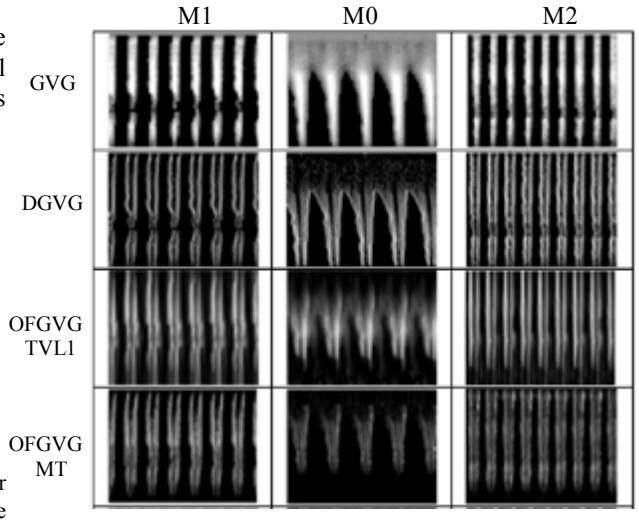


Fig.2: Representation of four playbacks (GVG, DGVG, OFGVG-TVL1, OFGVG-MT) for six different phonatory cases where either voice quality (pressed, creaky or breathy) or laryngeal mechanisms (M0, M1 or M2) are varied.

Similarities between DGVG and OFGVG playbacks are evidenced in Fig.2, especially in the shape appearance. However, OFGVG looks more blurred since movements taken into account by OF are not located only at the glottis edges (as in the DGVG case), but also in the vocal-folds surface. Breathy and M0 phonations present a posterior glottal chink that can be observed on GVG playback. Such regions are represented in DGVG and OFGVG playbacks with black color, as a result of the absence of movement. In creaky voice, OFGVG-TVL1 fails to provide accurate glottal cycles. The resulting playback may be improved by tuning the parameters of the preprocessing step.

Some features observed in DGVG playback and considered as artefacts due to segmentation problems do not appear in OFGVG-TVL1 playback. In M1 sequence for instance, the presence of mucus on vocal folds induces the appearance of a glottis splitted in two parts after the segmentation process. This is reflected by black spots in the median part of the glottis on GVG and DGVG playbacks. In both OFGVG playbacks, the glottis is not artificially splitted into two regions, as the motion field is robust to the presence of mucus.

C. Glottal velocity: Glottal Optical-Flow waveform

The Glottal Optical-Flow Waveform (GOFW) is a 1D representation of the velocity. GOFW is based on the same principle of the Glottal Area Waveform (GAW). The total magnitude of velocity is computed over the ROI for each instant of time. Graphically the GOFW represents the change of velocity as a function of time.

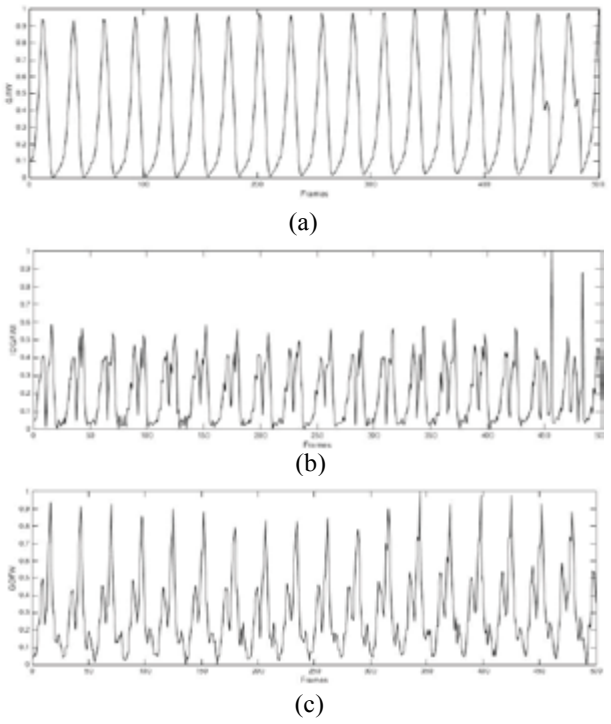


Fig.3: (a) GAW obtained by segmentation, (b) the absolute value of differentiated GAW (∠|DGAW∠|) and (c) GOFW.

The GOFW provide valuable information on the velocity instants over the HSV sequence. Additionally, if this information is overlapped with GAW one, it becomes feasible to analyze velocity variation with respect to glottal opening function. For instance, when maximal glottal opening is reached, GOFW shows a local minimum. Another interesting feature is that maximum speed is located during the closing phase. Since GOFW computes an absolute velocity, it is possible to obtain a similar representation by differentiating GAW and computing its absolute value (∠|DGAW∠|). As shown in Fig.3b, GOFW pulse shapes are similar to ∠|DGAW∠|, however with a stronger OF velocity during glottal closing.

IV. Conclusions and Discussion

High-speed videoendoscopy is probably the most promising technique for direct investigation of glottal dynamics in speech and singing. We have presented here a new approach to synthesize dynamical information from HSV recordings in a compact way, which does not depend on prior glottal segmentation. The glottis is treated as an unidentified object, and attention is focused on the motion field produced by vocal-folds vibration. Dense optical flow is computed among consecutive frames to extract dynamical information related to the pattern of glottal displacement. Three new playbacks are proposed to visualize the computed optical flow: OFKG, OFGVG and GOFW playbacks. There are some similarities in the information extracted from segmentation and OF, since both methods quantify the motion. However, the motion obtained from OF is raw information that include direction and magnitude of the pixels movements (displacement field map). Also using the displacement field is possible to segment the glottal gap, compute the contact time of the vocal folds and many other features. For the purpose of clinical diagnosis it seems a promising approach to complement, and eventually to replace, segmentation-based techniques.

Acknowledgements

This work has been funded by the Spanish Ministry of Economy and Competitivity under grant TEC2012 38630-C04-01.

References

[1] K. Kendall and R. Leonard, *Laryngeal Evaluation: Indirect Laryngoscopy to High-speed Digital Imaging*. Thieme Publishers Series. Thieme, 2010.

[2] J. G. Svec and H. K. Schutte, "*Videokymography: high-speed line scanning of vocal fold vibration,*" J. Voice, vol. 10, no. 2, pp. 201–5, Jun. 1996.

[3] D. D. Deliyski, P. P. Petrushev, H. S. Bonilha, T. T. Gerlach, B. Martin-Harris, and R. E. Hillman, "*Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution,*" Folia-Phoniatr. Logo, vol. 60, no. 1, pp. 33–44, Jan. 2008.

[4] J. Lohscheller and U. Eysholdt, "*Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics,*" IEEE Trans. Med. Imaging, vol. 27, no. 3, pp. 300–309, 2008.

[5] S-Z Karakozoglou, N. Henrich, C. d'Alessandro, and Y. Stylianou, "*Automatic glottal segmentation using local-based active contours and application to glottovibrography,*" Speech Communication, vol. 54, no. 5, pp. 641–654, 2011.

[6] B. Roubeau, N. Henrich, and M. Castellengo, "*Laryngeal vibratory mechanisms: The notion of vocal register revisited,*" Journal of Voice, vol. 23, no. 4, pp. 425 – 438, 2009.

[7] J. Sánchez Pérez, E. Meinhardt-Llopis, and G. Facciolo, *"TV-L1 Optical Flow Estimation"*, Image Processing On Line, 3 (2013), pp. 137–150

[8] G. Farneback, "*Two-frame motion estimation based on polynomial expansion,"* in Proceedings of the 13th Scandinavian Conference on Image Analysis, ser. SCIA'03. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 363–370.